

# ユーザレビューを用いたスポット対応付け手法の 提案とその応用

潘 健太<sup>1,a)</sup> 北山 大輔<sup>1,b)</sup>

受付日 2016年3月4日, 再受付日 2015年7月16日 / 2015年11月20日,  
採録日 2016年8月1日

**概要:** 近年, ユーザは Web 上の観光情報を活用して旅行計画を立てることが多くなっている. しかし, 旅行は未訪問スポットに行くことが多いため, 観光情報を適切に入手することは困難である. また, ユーザの未知なスポットに対する理解を支援するために, 訪問したことがある観光スポットの特徴を用いて, 未訪問エリアの観光スポットを説明する方法を提案する. 本研究では, まず, 観光スポットのユーザレビューを用いて特徴ベクトルを作成する. 次に, ユーザにとっての観光スポットの位置付けを抽出するためにあるスポットに対し, 既に訪れたスポットと比較した相対的特徴ベクトルを算出する. 最後に, 相対的特徴ベクトル間の類似度に基づいて既訪問スポットと未訪問スポットを関連付けし, その関係性を説明するためのキーワードを抽出する. また, プロトタイプシステムを構築し, 既訪問スポットと未訪問スポットとの説明情報の効果を検証する評価実験を行う.

**キーワード:** 観光スポット, 理解支援, ユーザレビュー, 分散表現

## Proposal of Spot Matching Method Using User Review And Its Application

HAN KENTA<sup>1,a)</sup> KITAYAMA DAISUKE HANAKO<sup>1,b)</sup>

Received: March 4, 2016, Revised: July 16, 2015/November 20, 2015,  
Accepted: August 1, 2016

**Abstract:** Most tourists resort to available information online when planning for leisure travel; however, this recourse often becomes problematic and misleading as tourist of the information may be directed to unfamiliar areas. On this regard, we proposed a method of explaining unfamiliar spots through the familiar features of spots they have visited. In this paper, at first, we generated the feature vector using user reviews of the tourist spot. Next, we used the relative feature vector compared with already visited spots to extract the role of the tourist spot for the user. Finally, we associated the visited spot with the unfamiliar spot by the similarity of the relative feature vector, and further extracted keywords that explain the relation. Furthermore, we developed a prototype of the system and evaluated the effect of the explanatory information between the familiar and unfamiliar spots.

**Keywords:** Tourist spots, explainability, tourist reviews, paragraph vector

### 1. はじめに

旅行先を決定するとき, 旅行者は観光スポット検索サイトや観光情報に関連する書籍を見て観光スポットを選び,

旅行計画を立てる. しかし, ユーザにとって訪問したいエリアは, 訪問したことがなく不慣れであることが多い. そのため, エリア内に数多く存在する観光スポットから, 自身の要求に合う観光スポットを見つけることは容易ではない.

<sup>1</sup> 工学院大学  
Kogakuin University

<sup>a)</sup> em18011@ns.kogakuin.ac.jp

<sup>b)</sup> kitayama@cc.kogakuin.ac.jp



図 1 プロトタイプシステムのユーザインタフェース

Tripadvisor<sup>\*1</sup>やじゃらん<sup>\*2</sup>などの観光スポット検索サイトでは、特定の観光スポットを訪れたことのあるユーザがレビューを投稿するため、観光スポットに関する豊富な情報が存在している。これらのレビューサイトは、ユーザの意思決定の材料になりえる。しかし、ユーザは検索エリアに関する事前知識がないため効率的に読むことは困難である。そこで我々は、さまざまな観光スポットを効果的に理解するためには、ユーザが訪問した経験のあるスポットを使って未訪問スポットと比較できることが重要であると考えた。例えば、日本の東京の「表参道」のような未訪問スポットは、パリでは「デシャンゼリゼ通り」と表現されていると、日本を初めて訪れるフランス人旅行者は「表参道」を理解しやすいであろう。この考え方は、ユーザが以前に経験した物事を現在の物事に適用する一種の類推である。既知の知識（ベースと呼ぶ）から概念（ターゲットと呼ぶ）を獲得するときに類推思考が働くと考えられる [1]。構造の類似性には 3 種類あり、特徴の共有数で決まる「対象レベルの類似性」、ベースに存在する関係とターゲットに存在する関係の共有度に基づく「関係レベルの類似性」、および題の解法あるいは目標レベルでの類似性である「プラグマティックな類似性」とがある [1], [2]。本研究では、類推の質を明示的に扱うため、構造の類似性「関係レベルの類似性」に近いと考えられる。

本研究では、ユーザの未知なスポットに対する理解を支援するため、既に訪問したことがある観光スポットの特徴を用いて、未訪問エリアの観光スポットを説明する手法を提案する。手法の詳細は 3 章で述べる。図 1 は、構築したプロトタイプシステムである。このように、抽出したキーワードを提示することで、ユーザの未訪問スポットに対する理解の支援を目指す。

## 2. 関連研究

ユーザの体験履歴を利用した検索や推薦システムに関する研究は数多く発表されている。倉島ら [3] は、Flickr に投稿された写真のジオタグ情報を人々の旅行履歴として利用した旅行ルート推薦手法を提案した。この手法では、ユーザの現在地から行きやすい場所とユーザの興味に合致し

た場所に移動しやすいと仮定し、行動モデルを生成している。ユーザのジオタグ付き写真集合は、時間情報でソートすると個人の旅行履歴とみなすことができると考え、ジオタグ情報を利用してユーザの行動モデルを生成している。Cheng らは [4]、自由に利用できるコミュニティ投稿の写真を活用して、パーソナライズされた旅行のおすすめに焦点を当て、特定のユーザプロフィールまたは属性を考慮し、パーソナライズされた旅行の推奨を行うことを提案した。類推は創造的思考に貢献すると指摘されてた [5]。

## 3. 未訪問スポットと既訪問スポットの対応付け

我々は、未訪問スポットを理解しやすくするために、もっとも位置付けの近いユーザの既訪問スポットを対応付け手法を提案する。まず、ユーザが既に訪問した複数の観光スポットとこれから訪問したい観光スポットエリア情報を入力する。本手法では、観光スポットのユーザレビューを用いて特徴ベクトルを生成する。次に、ユーザの観光地の位置付けを抽出するためあるスポットに対し、他の訪れたスポットと比較した相対的特徴ベクトルを算出する。同様に、未訪問スポットも同様にエリア内の各スポットの特徴ベクトルを求める。また、未訪問エリアの各観光スポットの相対的特徴ベクトルを、そのエリアの他の観光スポットと比較して計算する。次に、相対的特徴ベクトルの類似度によって既知スポットを未訪問スポットと対応付ける。最後に、その関係性を説明するためのキーワードを抽出する。

### 3.1 スポットのユーザレビューを用いた特徴ベクトル生成

本稿では、2016 年 9 月末までのじゃらんから得られたレビューデータを使用する。分散表現を用いて観光スポットの特徴ベクトルの作成する。このとき、観光スポット毎のレビューをまとめて 1 つの文書として扱う。本研究では、分散表現を計算するために Python のライブラリである gensim<sup>\*3</sup>を利用する。学習方法として、Distributed Bag-of-Words を利用して、各スポットの全レビューを使って 300 次元で作成したベクトルを使う。既訪問スポットや未訪問スポットのレビューベクトルは、形態素解析器である MeCab[6] に辞書「mecab-ipadic-NEologd」<sup>\*4</sup>を用いて、分かち書き（原型）したレビューを利用して作成する。相対的特徴ベクトル  $r_{state,i}$  を、式 1 として定義する。

$$r_{state,i} = s_i - average(S_{state} - s_i) \quad (1)$$

相対的特徴ベクトルは、そのスポット自体の特徴ベクトルから他のスポットの特徴ベクトルの平均を引いた値によって得られる。 $S_{state} = \{s_1, s_2, \dots, s_n\}$  は、既訪問ス

<sup>\*1</sup> <https://www.tripadvisor.com/>

<sup>\*2</sup> <https://www.jalan.net/kankou/>

<sup>\*3</sup> <https://radimrehurek.com/gensim/models/doc2vec.html>

<sup>\*4</sup> <https://github.com/neologd/mecab-ipadic-neologd/>

ポット集合もしくは未訪問スポット集合である． $state$  が  $f$  であれば，既訪問スポット集合を意味し， $state$  が  $u$  であれば，未訪問スポット集合を意味する． $s_i$  は集合  $S_{state}$  内の観光スポット  $i$  の特徴ベクトルを示している．

### 3.2 対応するスポットの決定

未訪問エリア内のスポットを既訪問スポットと対応付ける．そのため，未訪問スポットと既訪問スポットを，それらの相対的特徴ベクトルの類似度に基づいて対応付けを行う．類似度計算には，コサイン尺度を用いる．

関連付け手順について説明する．まず，ある未訪問スポットに最も類似度が高い既訪問スポットと対応付ける．ただし，類似度が閾値（本研究では 0.125）以下である場合は対応付けを行わない．

### 3.3 対応するスポットの特徴語の抽出

スポットの対応関係を示すだけでは，どういう点で対応するのかを理解するのは難しい．そこで，未訪問スポットと既訪問スポットの関係性を表すキーワードをユーザに提示する．しかし，分散表現である相対的特徴ベクトルから単語の特徴を得ることはできないので，他の方法を用いて単語を抽出する．

すべてのレビューは形態素解析器 MeCab によって単語を分割する．3.1 節の「mecab-ipadic-NEologd」辞書を利用する．ただし，助詞，助動詞，連体詞，記号，ストップワードを削除している．

キーワード抽出手順について説明する．まず，TFIDF 法を使って対象となる既訪問スポットと未訪問スポットの特徴語と TFIDF 値を求める．IDF 値を算出する集合は  $S_{state}$  を用いる．

2 つのスポットに共通する特徴語の TFIDF 値の調和平均を用いて，対応付けしたスポットの説明可能なキーワードとして抽出する．各単語のスコアを式 2 によって定義する．

$$score(t, d) = \frac{2 \times TFIDF(t, d, f) \times TFIDF(t, d, u)}{TFIDF(t, d, f) + TFIDF(t, d, u)} \quad (2)$$

$TFIDF(t, d, f)$  と  $TFIDF(t, d, u)$  は同じ単語に関する既訪問スポットにおける TFIDF 値と未訪問スポットにおける TFIDF 値を示している．単語スコアの上位  $N$  個の単語を説明情報としてユーザに提示する．

## 4. キーワード抽出に関する評価

### 4.1 実験内容

提案手法のキーワード抽出について評価を行うため，他のキーワード抽出手法と比較して評価を行う．以下の 3 つの方法を使って比較する．

#### A. 算術平均

表 1 キーワード説明情報における評価の割合

評価	方法 A	方法 B	方法 C
1	0.00%	0.00%	0.00%
2	28.28%	31.31%	29.29%
3	35.35%	31.31%	35.35%
4	10.10%	14.14%	12.12%
5	26.26%	23.23%	23.23%

#### B. 乗算

#### C. 提案手法 (調和平均)

方法 A と方法 B は，まず，各スポットの特徴として 3.1 節で作成された特徴ベクトルを使用する．そして，3.3 節を用いて既訪問スポットと未訪問スポットの特徴語と TFIDF 値を求める．

キーワードを抽出するとき，方法 A では，2 つのスポットに共通する特徴語のスコアとして算術平均を計算する．抽出した単語のスコアを式 3 によって定義する．

$$score(t, d) = \frac{TFIDF(t, d, f) + TFIDF(t, d, u)}{2} \quad (3)$$

方法 B では，2 つのスポットに共通する特徴語のスコアとして乗算を計算する．抽出した単語のスコアを式 4 によって定義する．

$$score(t, d) = TFIDF(t, d, f) \times TFIDF(t, d, u) \quad (4)$$

式 3，式 4 の  $TFIDF(t, d, f)$  と  $TFIDF(t, d, u)$  は同じ単語に関する既訪問スポットの TFIDF 値と未訪問スポットの TFIDF 値を示している．単語スコアの上位  $N$  個の単語を説明情報としてユーザに提示する．

### 4.2 実験手順

クラウドソーシングのサービスである，CrowdWorks<sup>\*5</sup>を利用して 23 人の被験者を集めた．各方法を用いて，被験者の既訪問スポットに基づいて，未訪問スポットの説明可能な情報を提示した．まず，被験者は自分が既に訪れたことのある観光スポットを入力した．被験者は 4 から 10 個のスポットを入力した．本実験では，入力文字列とシステム内のスポット名を照会するために，入力文字列から検索したスポット名候補の中から対象スポットを選択した．次に，被験者は旅行などに行ったことがなく，これから訪れたい都道府県やエリアを入力した．

本手法では，方法 A から C それぞれにおいて，関連付けられた既訪問スポットと未訪問スポット，およびスポット同士の関係を説明するキーワードを提示した．キーワード ( $N \leq 5$ ) は最大 5 つを提示した．結果を評価するために，被験者は以下の 5 つの選択肢から 1 つを選んだ．

(1) キーワードなし．

(2) 2 つのスポットには関係性があり，キーワードによって関係が明確になった．

<sup>\*5</sup> <https://crowdworks.jp/>

- (3) 2つのスポットの関係性があり、キーワードによって初めて気がついた。
- (4) 2つのスポットには関係性があるが、キーワードは関係を表していない。
- (5) 2つのスポットには関係性がない。

### 4.3 実験結果と考察

表1は、方法AからCにおける選択肢1から5それぞれの選択割合を示している。方法Bでは、多くの被験者が選択肢2を選択している。このことから、方法Bは明らかな関係を表現するキーワード抽出ができることがわかった。方法AとCでは、選択肢2の割合が減少し、選択肢3の割合が増加している。このことから、方法AとCは隠れた関係を表現するキーワード見つけて被験者に提示することができる考えた。しかし、方法Aでは、選択肢5の割合が増加しているため、方法Cより無関係なキーワードを抽出しやすくなったといえる。また、選択肢2と3を合計すると、方法Cが最も高くなった。このことから、説明するためのスポットの関連を説明するキーワードの精度が最も高いといえる。

## 5. 対応付けに関する評価

### 5.1 実験内容

提案手法を他の手法と比較して評価を行う。以下の3つの方法を使って比較する。

- D. メタデータ (カテゴリ, 滞在時間, 訪問時期)
- E. 分散表現 (特徴ベクトル)
- C. 提案手法 (相対的特徴ベクトル)

方法Dは、観光スポット検索サイトでスポットを検索するためのメタデータである。次のように、検索によく使われる3つのメタデータを選択する。

- カテゴリ: 神社・神宮・寺院, 観光施設・名所巡り等
- 滞在時間: 1時間未満, 1-2時間等
- 訪問時期: 1-12月, 春, 夏, 秋, 冬

方法Dでは、同じカテゴリ, 同じ滞在時間, 同じ季節の既訪問スポットと未訪問スポットを抽出する。抽出できない場合は、季節, 滞在時間, カテゴリの順に情報を削除している。未訪問スポットが複数ある場合は、レビュー数が最も多いスポットを選択する。最後に、3.3節で説明した方法により説明可能な単語を抽出し、被験者に提示する。

方法Eでは、各スポットの特徴として3.1節で作成された特徴ベクトルを使用する。そして、3.3節を用いて説明可能な単語を抽出し、被験者に提示する。

クラウドソーシングのサービスである、CrowdWorksを利用して24人の被験者を集めた。各方法を用いて、被験者の既訪問スポットに基づいて、未訪問スポットの説明可能な情報を提示した。実験手順は4.2節で説明した手順と同じである。

### 5.2 実験結果と考察

使用可能なデータの合計は285である。表2は、それぞれの方法における選択肢1から5それぞれの選択割合を示している。方法Dについて、多くの被験者が選択肢2を選択している。このことから、方法Dは明らかな関係にあるスポット同士を関連付けることができることがわかった。他の方法と比較して、方法C(提案手法)では、選択肢2の割合が減少し、選択肢3の割合が増加している。このことから、方法Cは隠れた関係を見つけて被験者に提示することができる考えた。しかし、選択肢5の割合も増加しているため、無関係なスポットを抽出しやすくなったといえる。

選択肢2と3を合計すると、方法Dが最も高くなる。このことから、説明するためのスポットの関連付けの精度が最も高いといえる。しかし、方法Dは対応付けすることができるスポットの数が少数であり、未訪問スポットの多くに説明を付与することができない。

我々は、選択肢2と3はトレードオフの関係にあると考えている。選択肢2は説明する必要がない関係であり、選択肢3は説明に値する関係である。選択肢2を維持しながら、選択肢3をできるだけ増やすことが重要であると考ええる。この観点から、方法Eと提案手法は同程度の精度を示していると考えられる。

選択肢3は、キーワードの抽出に失敗した場合に選択肢5になる可能性が高いと考えられる。したがって、キーワード抽出方法の改善が必要である。

方法Eと方法Cでは、被験者が入力した既訪問スポットのカテゴリが異なる場合は同様の傾向を示した。既訪問スポット集合では、同一カテゴリに含むスポットが半数を超えた場合はカテゴリは同じと定義する。それ以外の場合はカテゴリは異なると定義する。表2は選択肢2と選択肢3の評価の割合である。提案手法である方法Cを使用すると、被験者が既訪問スポットと関係なしで、被験者に有用なキーワードを提示することができる。相対的特徴ベクトルを用いることによって、カテゴリをこえて各スポットの特徴を求めることができるといえる。既訪問スポットのジャンルが類似する場合は、方法Eの方がより良い評価となっている。入力スポットの種類によって対応付け手法を切り変えることが有効と考えられる。

## 6. 未訪問スポット説明の有効性評価

### 6.1 実験内容

提案手法の有効性を評価するために以下の2つのシステムを使って比較する。

- a. 未訪問スポット名のみを表示
- b. 未訪問スポット名, 関連する入力スポット, 関連を説明するキーワードを表示 (提案手法)

被験者は4個から10個の自分が既に訪れたことのある

表 2 対応付け説明情報における評価の割合

評価	方法 D	方法 E	方法 C
1	0.00%	0.00%	0.00%
2	41.30%	33.85%	29.36%
3	43.48%	47.69%	48.62%
4	2.17%	2.31%	2.75%
5	13.04%	16.15%	19.27%

表 3 既訪問スポットカテゴリが異なる場合と同じ場合の評価割合

	異なる場合	同じ場合
説明情報 E&評価 2	56.82%	43.18%
説明情報 C&評価 2	71.87%	28.13%
説明情報 E&評価 3	51.61%	48.39%
説明情報 C&評価 3	52.83%	47.17%

観光スポットを入力した。本実験では、入力文字列とシステム内のスポット名を照会するために、入力文字列から検索したスポット名候補の中から対象スポットを選択した。次に、被験者は旅行などに行ったことがなく、これから訪れたい異なる 2 つの都道府県やエリアを入力した。これに対し、システム a とシステム b はランダムな順番で実行される。システムを評価するために、被験者は以下の 2 つの設問について回答した。また、それらの選択理由について自由記述で答えた。

- (1) 表示されたスポットの詳細情報はどちらの方が分かりやすいかを選択してください。
- (2) 旅行の計画を立てる際にどちらの方が使いたいかを選択してください。

## 6.2 実験結果と考察

スポットの詳細情報はどちらの方が分かりやすいかに対する回答について、システム a とシステム b の結果はそれぞれ 12 件と 38 件となった。被験者の回答理由の例を表 4(システム a) および、表 5(システム b) に示す。また、旅行の計画を立てる際にどちらの方が使いたいかに対する回答について、システム a とシステム b の結果はそれぞれ 10 件と 40 件となった。被験者の回答より、得られた 2 つのシステムの特徴をまとめる。

システム b では、キーワードや関連している情報が表示されており分かりやすいという特徴がある。これに対し、システム a では、スポット名のみが表示されているためシンプルで分かりやすいという特徴があるが、なぜそのスポットが検索されたのか分からないなどの否定的な回答も多数ある。システムの表示情報が「シンプルな方が良い」と「詳細な情報が良い」の 2 つに大きく分けられているが、80%の被験者はシステム b の方が使いたいと回答しており、提案手法の方が良い結果であるといえる。また、システム b に関してスポットの写真を載せるとより分かりやすいなどの意見があり、画像による関連性の説明という拡張が考えられる。

表 4 被験者がシステム a の方が良いと回答した例

スポット名だけ書いてあるのでごちゃごちゃにならず、一目でどこに何があるか分かるからです。
スポットだけ表示されるところがシンプルで分かりやすかったから。
シンプルにそのエリアの観光スポットを知ることが出来る。

表 5 被験者がシステム b の方が良いと回答した例

自分が訪れた事のあるスポットと関連性があり連想しやすいから。
どのような場所か想像しやすいので、計画を立てるのにも使いたいと思ったから。
何と関連して表示された等詳細の情報が表示されわかりやすかった。

## 7. まとめと今後の課題

本研究では、ユーザが行きたい観光スポットが決まっている場合、ユーザの事前知識が不足しているため、観光検索サイトを使用してランキング、おすすめ情報やカテゴリなどを基に検索された観光スポットに対する理解が困難であることに着目した。未訪問スポットに対する理解を支援するために、未訪問スポットをユーザが既に訪れたことがある既訪問スポットと比較することによって、理解を支援する説明手法を提案した。

キーワード抽出評価について、提案手法の方が既訪問スポットによる未訪問スポットの説明性が高いことを確認した。対応付け手法の結果として、カテゴリを利用する場合では、未訪問スポットと関係する既訪問スポット最も少なくなることがわかった。相対的特徴ベクトルと調和平均を利用することによって、各スポットの特徴を求めることができた。また、未訪問スポットと既訪問スポットに対して意外性がある関連付けが可能で、ユーザが知らない観光スポットに対する興味と関心を集めることができる可能性があることを確認した。システム評価について、提案手法の方がより詳細な情報を提示しているため、ユーザが未訪問スポットをイメージしやすいことを確認した。

今後の課題としては、レビュー数が少ないスポットに対する関連付けの精度が低いと、新たな方法を検討する必要がある。また、ある観光スポットに類似する特徴を持つ既訪問スポットで説明するだけでなく、既訪問スポット集合と対象観光スポットの関係を説明可能な可視化インタフェースへと拡張する予定である。

謝辞 本研究の一部は、平成 30 年度科研費基盤研究(C)(課題番号：18K11551)によるものです。ここに記して謝意を表すものとします。

## 参考文献

- [1] D. Gentner, "Structure-Mapping: A Theoretical Framework for Analogy", Cognitive Science, Vol.7, pp.155-170, 1983
- [2] K. J. Holyoak and P. Thagard, "Analogical Mapping by Constraint Satisfaction", Cognitive Science, Vol.13, pp.295-355, 1989

- [3] T. Kurashima, T. Iwata, G. Irie and K. Fujimura., “Travel route recommendation using geotags in photo sharing sites”, CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, pp.579-588, 2010
- [4] A. J. Cheng, Y. Y. Chen, Y. T. Huang and Winston H. Hsu, “Personalized Travel Recommendation by Mining People Attributes from Community-Contributed Photos”, MM '11 Proceedings of the 19th ACM international conference on Multimedia, pp.83-92, 2011
- [5] K. J. Holyoak and P. Thagard, “Mental Leaps: Analogy in Creative Thought, MIT Press”, Journal of Japanese Society for Artificial Intelligence, Vol.11, No.3, pp.489, 1996
- [6] T. Kudo, K. Yamamoto and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004



### 潘 健太

1996 年生。2018 年工学院大学情報学部コンピュータ科学科卒業。2018 年同大学大学院工学研究科情報専攻修士課程入学，現在に至る。観光スポット検索に興味を持つ。



### 北山 大輔

2009 年兵庫県立大学大学院環境人間学研究科博士後期課程修了。同年同学科客員研究員。2011 同学科特任助教。2012 年工学院大学情報学部助教。2016 年同学部准教授，現在に至る。博士(環境人間学)。Web マイニング，地理情報検索/推薦ベースを研究。電子情報通信学会。日本データベース学会各会員。